<<使用C#开发搜索引擎>>

图书基本信息

书名：<<使用C#开发搜索引擎>>

13位ISBN编号：9787121633973

10位ISBN编号：7121633973

出版时间：2011-11-18

出版时间：清华大学出版社

作者：罗刚

<<使用C#开发搜索引擎>>

内容概要

介绍如何以C#作为工具开发搜索引擎。

全书以完成一个网站搜索\垂直搜索作为目标。
从网络爬虫抓取数据开始，然后到中文分词、文本排重等文本挖掘技术和搜索结果展现。
本书是唯一介绍业界热门的Lucene.Net、使用WebBrowser做爬虫以及结合Solr开发ASP.NET搜索的书籍。
从C#基础开始，逐渐深入，是学习搜索引擎开发的首选。

对于学习复杂数据结构和应用动态规划等常用算法也有参考价值。

作者简介

罗刚,猎兔搜索(http://www.lietu.com)创始人。
创建包括旅游搜索和舆情监测在内的多个技术开发团队。
有多年软件培训经验，相关学员已经在京东商城、UCWeb、MadeInChina等多家公司从事技术开发。
出版过的相关书籍包括《自己动手写搜索引擎》、《自己动手写网络爬虫》、《解密搜索引擎技术实战》以及视频教程《Lucene构建网站搜索系统》。

<<使用C#开发搜索引擎>>

书籍目录

使用C#开发搜索引擎 1
第1章 使用C#开发搜索引擎快速入门 2
1.1 各种搜索引擎 2
1.1.1 通用搜索 2
1.1.2 垂直搜索 3
1.1.3 站内搜索 4
1.2 搜索引擎整体结构 4
1.3 搜索引擎基本技术 5
1.3.1 网络爬虫 5
1.3.2 文本挖掘 6
1.3.3 全文索引 6
1.3.4 搜索语法介绍 10
1.3.5 搜索用户界面 11
1.4 C#开发快速入门 13
1.4.1 准备开发环境 13
1.4.2 基本语法 13
1.4.3 多维数组 15
1.4.4 位运算 15
1.4.5 枚举类型 16
1.4.6 面向对象 17
1.4.7 集合类 19
1.4.8 泛型 21
1.4.9 委托和事件 21
1.4.10 类库 24
1.5 本章小结 24
1.6 术语表 25
第2章 使用C#开发网络爬虫 26
2.1 网络爬虫抓取原理 26
2.2 爬虫架构 29
2.2.1 基本架构 29
2.2.2 分布式爬虫架构 31
2.2.3 垂直爬虫架构 32
2.3 下载网页 33
2.3.1 HTTP协议 33
2.3.2 下载静态网页 37
2.3.3 下载动态网页 41
2.4 网络爬虫遍历与实现 49
2.5 网站地图 51
2.6 连接池 52
2.7 URL地址查新 53
2.7.1 嵌入式数据库 54
2.7.2 布隆过滤器 56
2.8 抓取RSS 59
2.9 解析相对地址 61
2.10 网页更新 62

2.11 信息过滤 64
2.12 垂直行业抓取 70
2.13 抓取限制应对方法 70
2.13.1 更换IP地址 70
2.13.2 抓取需要登陆的网页 73
2.13.3 抓取ASP.NET网页 76
2.14 保存信息 79
2.14.1 存入数据库 79
2.14.2 存成图像 80
2.15 日志 81
2.16 本章小结 84
2.17 术语表 85
第3章 索引各种格式文档 89
3.1 从HTML文件中提取信息 89
3.1.1 识别网页的编码 89
3.1.2 正则表达式 91
3.1.3 Html Agility Pack介绍 96
3.1.4 网页正文提取 100
3.1.5 结构化信息提取 113
3.1.6 查看网页的DOM结构 117
3.1.7 网页结构相似度计算 119
3.2 从非HTML文件中提取文本 122
3.2.1 TEXT文件 122
3.2.2 PDF文件 123
3.2.3 Office文件 125
3.2.4 Rtf文件 126
3.3 本章小结 128
3.4 术语表 128
第4章 自然语言处理 129
4.1 统计机器学习 129
4.1.1 协同推荐 130
4.2 文档排重 136
4.3 中文关键词提取 145
4.3.1 关键词提取的基本方法 146
4.3.2 从网页中提取关键词 149
4.4 相关搜索 149
4.5 拼写检查 150
4.5.1 拼写检查的概率模型 151
4.5.2 模糊匹配问题 152
4.5.3 英文拼写检查 156
4.5.4 中文拼写检查 159
4.6 文本摘要 160
4.6.1 文本摘要的设计 160
4.6.2 实现文本摘要技术 161
4.6.3 Lucene.Net中的动态摘要 167
4.7 文本分类 168
4.7.1 自动分类的接口定义 168

4.7.2 自动分类的实现 169
4.8 自动聚类 170
4.8.1 文档相似度 171
4.8.2 K均值聚类方法 174
4.8.3 K均值实现 176
4.9 拼音转换 178
4.10 句法分析树 178
4.11 信息提取 187
4.12 本章小结 194
4.13 术语表 196
第5章 用C#实现中文分词 197
5.1 汉语中的词 197
5.2 文本切分的基本方法 197
5.3 有限状态机 199
5.4 查找词典算法 201
5.4.1 标准Trie树 202
5.4.2 三叉Trie树 208
5.5 中文分词的原理 213
5.6 中文分词流程与结构 217
5.7 切分词图 219
5.7.1 保存切分词图 220
5.7.2 生成全切分词图 224
5.8 概率语言模型的分词方法 227
5.8.1 一元模型 228
5.8.2 N元模型 231
5.9 最大熵 237
5.10 未登录词识别 238
5.11 词性标注 239
5.12 地名切分 252
5.12.1 地址类性标注 252
5.12.2 未登录词识别 253
5.13 本章小结 254
5.14 术语表 255
第6章 Lucene.Net原理与应用 256
6.1 Lucene.Net快速入门 256
6.1.1 索引文档 257
6.1.2 搜索文档 258
6.1.3 Lucene.Net结构 260
6.2 Lucene.Net深入介绍 260
6.2.1 索引原理 261
6.2.2 分析文本 263
6.2.3 遍历索引库 267
6.2.4 检索模型 268
6.2.5 收集最相关的文档 270
6.3 索引中的压缩算法 275
6.3.1 变长压缩 276
6.3.2 差分编码 278

6.4 创建和维护索引库 278
6.4.1 设计一个简单的索引库 279
6.4.2 创建索引库 280
6.4.3 向索引库中添加索引文档 280
6.4.4 删除索引库中的索引文档 283
6.4.5 更新索引库中的索引文档 284
6.4.6 索引的优化与合并 284
6.5 查找索引库 285
6.5.1 布尔查询 286
6.5.2 同时查询多列 289
6.5.3 跨度查询 290
6.5.4 通配符查询 294
6.5.5 过滤 294
6.5.6 按指定列排序 295
6.5.7 查询大容量索引 300
6.5.8 函数查询 302
6.5.9 定制相似性 305
6.5.10 评价搜索结果 307
6.6 中文信息检索 308
6.6.1 Lucene.Net中的中文处理 308
6.6.2 Lietu中文分词的使用 309
6.6.3 定制Tokenizer 310
6.6.4 解析查询串 312
6.6.5 实现字词混合索引 315
6.7 抓取数据库中的内容 319
6.7.1 读取数据 319
6.7.2 数据同步 321
6.8 概念搜索 321
6.9 本章小结 324
6.10 术语表 325
第7章 实现搜索用户界面 327
7.1 搜索页面设计 327
7.1.1 用于显示搜索结果的ASP.NET 327
7.1.2 搜索结果分页 330
7.1.3 设计一个简单的搜索页面 331
7.2 实现搜索接口 331
7.2.1 Lucene.Net搜索接口 331
7.2.2 指定范围搜索 336
7.2.3 搜索页面的索引缓存与更新 337
7.3 实现关键词高亮显示 340
7.4 实现分类统计视图 341
7.4.1 搜索结果分类统计与导航 341
7.4.2 层次树 345
7.5 相关搜索词 348
7.6 实现AJAX自动完成 349
7.6.1 总体结构 349
7.6.2 服务器端处理 350

7.6.3 浏览器端处理 350
7.7 集成其他功能 353
7.7.1 拼写检查 353
7.7.2 再次查找 353
7.7.3 黑名单 354
7.7.4 搜索日志 355
7.8 本章小结 356
第8章 使用Solr开发网站搜索 357
8.1 搜索服务器端 357
8.1.1 Solr结构 358
8.1.2 启动Solr服务器 359
8.1.3 配置支持中文的Solr 362
8.1.4 索引数据 366
8.1.5 查询功能 367
8.1.6 高亮 370
8.2 Solr的.NET客户端 371
8.2.1 使用SolrNet 372
8.2.2 实现多分类 380
8.3 查询语法 382
8.3.1 对空格的支持 382
8.3.2 日期加权 382
8.4 索引分布 385
8.5 本章小结 387

版权说明